Received: 31 July 2017 | Accepted: 22 September 2017

DOI: 10.1111/2041-210X.12910

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing

Thierry Chambert[1,2] | J. Hardin Waddle[3] | David A. W. Miller[1] | Susan C. Walls[4] | James D. Nichols[2]

[1]Department of Ecosystem Science and Management, Pennsylvania State University, University Park, PA, USA

[2]Patuxent Wildlife Research Center, U.S. Geological Survey, Laurel, MD, USA

[3]Wetland and Aquatic Research Center, U.S. Geological Survey, Lafayette, LA, USA

[4]Wetland and Aquatic Research Center, U.S. Geological Survey, Gainesville, FL, USA

Correspondence
Thierry Chambert
Email: thierry.chambert@gmail.com

Funding information
United States Geological Survey's Greater Everglades Priority Ecosystem Science programme (USGS GEPES); Amphibian Research and Monitoring Initiative (USGS ARMI)

Handling Editor: Nigel Yoccoz

Abstract

1. The development and use of automated species detection technologies, such as acoustic recorders, for monitoring wildlife are rapidly expanding. Automated classification algorithms provide cost- and time-effective means to process information-rich data, but often at the cost of additional detection errors. Appropriate methods are necessary to analyse such data while dealing with the different types of detection errors.
2. We developed a hierarchical modelling framework for estimating species occupancy from automated species detection data. We explore design and optimization of data post-processing procedures to account for detection errors and generate accurate estimates. Our proposed method accounts for both imperfect detection and false-positive errors and utilizes information about both occurrence and abundance of detections to improve estimation.
3. Using simulations, we show that our method provides much more accurate estimates than models ignoring the abundance of detections. The same findings are reached when we apply the methods to two real datasets on North American frogs surveyed with acoustic recorders.
4. When false positives occur, estimator accuracy can be improved when a subset of detections produced by the classification algorithm is post-validated by a human observer. We use simulations to investigate the relationship between accuracy and effort spent on post-validation, and found that very accurate occupancy estimates can be obtained with as little as 1% of data being validated.
5. Automated monitoring of wildlife provides opportunity and challenges. Our methods for analysing automated species detection data help to meet key challenges unique to these data and will prove useful for many wildlife monitoring programmes.

KEYWORDS

acoustic recorder units, detection optimization, false positives, *Hyla cinerea*, information-rich data, *Lithobates grylio*, sensitivity, signal processing, specificity

1 | INTRODUCTION

Acoustic recording units (ARU) and other automated species detection methods, such as camera traps (O'Connell, Nichols, & Karanth, 2010), have emerged as prominent tools for monitoring wildlife populations (Acevedo & Villanueva-Rivera, 2006; Blumstein et al., 2011; Dorcas, Price, Walls, & Barichivich, 2010; Wrege, Rowland, Keen, & Shiu, 2017). Automated recording devices can capture large numbers of animal "detections" (e.g. vocalizations, photos) for many species across extended windows of time. These recorders allow researchers to monitor species that are often difficult to detect under standard protocols and provide an economical alternative when making many repeat visits to a site is cost prohibitive (Zwart, Baker, McGowan, & Whittingham, 2014). Recordings also provide a digital archive that can be revisited as new questions and methods arise. This utility has spurred the popularity of ARUs, and the technology is being applied to a wide range of species including bats (MacSwiney, Cristina, Clarke, & Racey, 2008), whales (Clark & Clapham, 2004), elephants (Wrege et al., 2017), birds (Haselmayer & Quinn, 2000; Rempel, Hobson, Holborn, Van Wilgenburg, & Elliott, 2005) and anurans (Crouch & Paton, 2002). Acoustic recorders are also being used for long-term monitoring of entire ecosystems (Lammers, Brainard, Au, Mooney, & Wong, 2008). A major challenge encountered when collecting ARU data is the need to efficiently process and analyse the large quantity of data produced by these devices (Wrege et al., 2017).

Acoustic recording units can be used to characterize species' distribution across space and time by estimating occurrence probabilities (i.e. occupancy). Understanding spatial variation in occurrence can help characterize habitat relationships and patterns of species co-occurrence (Miller, Brehme, Hines, Nichols, & Fisher, 2012), whereas temporal variation may be analysed to assess meta-population dynamics (MacKenzie, Nichols, Hines, Knutson, & Franklin, 2003) or provide insights into phenology (Chambert, Kendall, et al., 2015). The volume of data potentially collected by ARU presents a significant challenge in processing the observations for analysis. Consequently, computational algorithms to recognize and classify calls are now commonly employed to identify species (Brandes, 2008). Much work has gone into developing reliable algorithms (e.g. Bardeli et al., 2010). However, even in the best case scenario, sensitivity (the probability a species call is correctly detected) and specificity (the probability that calls of other species or other sounds [non-calls] are not misidentified as calls of the target species) of algorithm-based methods will be less than one (Acevedo, Corrada-Bravo, Corrada-Bravo, Villanueva-Rivera, & Aide, 2009). Increasing sensitivity of processing algorithms generally comes at a cost of lower specificity (i.e. higher rates of misidentification), and this latter error type poses a significant challenge when trying to estimate patterns of occurrence (McClintock, Bailey, Pollock, & Simons, 2010; Miller, Weir, et al., 2012). Even when they occur at low probabilities, misidentifications (i.e. false positives), will quickly accumulate for methods such as acoustic recording due to the volume of observations (Simons, Alldredge, Pollock, & Wettroth, 2007). Methods that account for detection uncertainty can be highly biased if they do not explicitly deal with false-positive errors (Miller et al., 2015), again even if those error rates are low. Thus, robust methods for estimating spatial and temporal occurrence patterns with ARU data should account for both imperfect detection and false-positive errors.

Recently developed approaches for dealing with misidentification when estimating species occupancy provide a useful starting point for the analysis of data obtained from automated devices such as ARUs (Chambert, Miller, & Nichols, 2015; Miller et al., 2011, 2013; Royle & Link, 2006). These approaches account for situations where a detection may occur for sites that are unoccupied by the species of interest (i.e. false-positive detection) and thus provide unbiased estimators for occupancy. This is accomplished using additional information that can help to distinguish true and false-positive detections. Usually, such additional information consists of a subset of unambiguous detections, i.e. detections that are confirmed as being true positives, without uncertainty. For example Miller et al. (2011) proposed an approach that combines data collected with methods that do and do not lead to false-positive detections. Alternatively, Chambert, Miller et al., (2015) outlined an approach where a subset of observations is *a posteriori* verified to determine the proportion of true and false-positive detections. Existing approaches, however, do not utilize any information on the number of detections (when >1) recorded at each survey occasion. They do not account explicitly for cases where multiple detections may be recorded within a sampling occasion; for instance when multiple call events are recorded by an ARU or multiple scats are found during a visual encounter survey. Instead all detections are traditionally combined from the same survey into a single detection-event (i.e. simply coded as a "1") to apply a binomial model (Royle & Link, 2006). This inefficiently uses information available in the data. The number of detection events that occurs within a single survey occasion provides useful information when trying to disentangle true detections from false positives. One would expect that in most cases more detections will occur at an occupied site (Chambert, Miller et al., 2015).

We develop a new occupancy modelling approach that accounts for false negatives and false positives and efficiently uses information about the frequency of detections to improve estimates. We assess performance of three models, a general model (*G*) and two simpler versions (*S1* and *S2*) using simulations and real datasets. Here, our focus is primarily on ARU data, but the approach can be used for any type of datasets that consists of multiple detections, potentially including misidentification. We discuss our findings to provide guidance and help future investigators optimize data processing.

2 | MODEL DESCRIPTION

Consider the case for data generated by an automatic recording device (e.g. ARU) where (1) recordings are processed using an automated classification algorithm to identify call events for a focal species and (2) a subset of these algorithm-generated detections is verified (i.e. classified as being either true or false) independently. Recordings are

obtained from ARU's deployed at a fixed number of sites $i = 1, \ldots I$ and for a given time period. Typically recordings are taken for short periods at regular intervals (e.g. 5 min every hour) during the day or night to match peak acoustic activity of the focal species (Walls et al., 2014). This pattern is often repeated over a prolonged period (e.g. several weeks or months). As a result, the number of individual recording sessions may be very large, and it will often be preferable to combine several recording sessions (e.g. several days of recording) into a single sampling session to save computation time. These occasions are denoted as $j = 1, \ldots J$. For example consider the case where recorders were active for 5 min per hour during six consecutive hours each night and this was repeated for a period of 12 weeks. We can bin data so that each week is a sampling occasion, where each occasion j corresponds to 210 min of recording (i.e. 7 nights × 30 min per night). The number of sampling occasions will depend on sampling protocol and field constraints, and it is important to consider the assumption of independence, across consecutive occasions, when adjusting the sampling window.

For each site i and occasion j, we will thus end up with a number M_{ij} of algorithm-generated detections for the species of interest. The variable M_{ij} is a non-negative integer, and can consist of a combination of true positives (i.e. calls from the focal species) and false positives (i.e. misidentified background noise or calls from another species). For model *G*, we use two data summaries obtained from M_{ij}: (1) binary detection data y_{ij}, where $y_{ij} = 1$ if $M_{ij} > 0$ and $y_{ij} = 0$ if $M_{ij} = 0$; and (2) the total number N_i of detections for a site i, where $N_i = \sum_{j=1}^{J} M_{ij}$. We also use data obtained from the validation (i.e. confirmation as true or false positive) of a subset of these detections (see below). The alternative model *S1* only uses binary detection data, y_{ij}, whereas model *S2* only uses the count data N_i. We will first describe model *G*, and then, we will briefly describe models *S1* and *S2* as particular cases of *G*.

Model *G* can be broken down, hierarchically, into four components: (1) the true ecological occupancy process; (2) the longitudinal binary detection process; (3) the detection-count process for the total number of true and false positives and (4) the partial verification process that partitions a subset of detections into true and false categories. Site occupancy is denoted by the binary latent variable z_i, which takes value $z_i = 1$ if a site is occupied and $z_i = 0$ if it is unoccupied. The probability $z_i = 1$ (i.e. the species is truly present) is given by ψ_i, where

$$z_i \sim \text{Bernoulli}(\psi_i)$$

We then specify a longitudinal binary detection process that accounts for the possibility of false negatives, i.e. cases when an occupied site gets no detection. False negatives arise from two different processes: (1) a species might be present at a site, but not calling during recording time; (2) a species might be present, calling and recorded by the ARU, but the algorithm failed to detect any call. At any sampling occasion j, a site i can either yield at least one detection ($y_{ij} = 1$ if $M_{ij} \geq 1$ detection) or not ($y_{ij} = 0$ if $M_{ij} = 0$). The outcome $y_{ij} = 1$ is conditional on the occupancy status of site i, and it occurs with probabilities p_{11} and p_{10} for occupied and unoccupied sites respectively:

$$y_{ij}|z_i \sim \text{Bernoulli}(p_i),$$

where $p_i = z_i * p_{11} + (1 - z_i) * p_{10}$

This part of the model also applies to model *S1*, but not to model *S2*. It is important to keep in mind that at unoccupied sites, any algorithm-generated detection is by definition a false positive, whereas at occupied sites, an individual detection can either be a true or a false detection. Therefore, probabilities p_{11} and p_{10} are defined as site-level (*sensu* Chambert, Miller, et al., 2015) probabilities of true and false positives, respectively, and can be written as $p_{11} = \Pr(y_{ij} = 1 | z_i = 1)$ and $p_{10} = \Pr(y_{ij} = 1 | z_i = 0)$. At this stage, the model is equivalent to the binomial mixture model first presented by Royle and Link (2006), and without extra information or assumption, false positives cannot be disentangled from true detections.

The next two steps of our model utilize the extra information provided by the total number of detections and the state verification of a subset of these detections (i.e. classification as true or false detections). Total numbers of true (K_i) and false (Q_i) detections for a site i ($N_i = K_i + Q_i$) are modelled with two distinct Poisson processes, with intensity rate parameters λ and ω, respectively, that are conditional on both z_i and y_{ij}. If at least one detection occurred at site i (i.e. $\sum_{j=1}^{J} y_{ij} > 0$), the sum $N_i = K_i + Q_i$ must be constrained as strictly positive. Because of this constraint, we thus use a Zero-Truncated Poisson (ZT Poisson) distribution to model N_i. Let us define a site-specific variable $w_i = \sum_{j=1}^{J_i} y_{ij}$. When $w_i = 0$, we have:

$$N_i = 0, K_i = 0 \quad \text{and} \quad Q_i = N_i - K_i = 0$$

and, when $w_i = 1$, we have:

$$N_i \sim \text{Poisson}(\lambda * z_i + \omega)$$

$$K_i \sim \text{Poisson}(\lambda * z_i)$$

$$Q_i = N_i - K_i$$

Finally, we deal with the process of *a posteriori* data validation: for each site i, a subset n_i ($n_i \in \{0; N_i\}$) of the N_i algorithm-generated detections are randomly chosen and confirmed, by a human observer, as true or false detections. The output of this validation process is the number of true positives (k_i) among the n_i validated data. Note that the number of false positives validated (q_i) is simply $n_i - k_i$. For model *G* and *S2*, this probabilistic process is described with a hypergeometric distribution: the k_i true positives are obtained from a random draw, without replacement, of n_i detections, within a finite population of size N_i (total number of detections at site i) that contains exactly K_i true detections. Therefore, we model data k_i, conditionally on the known values n_i and N_i, and the unknown variable K_i, as:

$$k_i \sim \text{Hypergeometric}(K_i, N_i, n_i)$$

For model *S1*, the information from data validation is simply incorporated as data into the model as known site occupancy status, $z_i = 1$, for sites at which at least one true detection ($k_i > 0$) was confirmed. Model *S1* consists of only the first two processes of model *G*, i.e. the occupancy process and the binomial detection process. It can thus be written as:

$$z_i \sim \text{Bernoulli}(\psi_i)$$

with

$$y_{ij}|z_i \sim \text{Bernoulli}(p_i),$$

$$\text{with } p_i = z_i * p_{11} + (1 - z_i) * p_{10}$$

Model *S2* consists of processes 1, 3 and 4 of model *G*. It does not include the binomial process describing y_{ij}, and can thus be expressed as:

$$z_i \sim \text{Bernoulli}(\psi_i)$$

$$K_i|z_i \sim \text{Poisson}(\lambda * z_i)$$

$$Q_i \sim \text{Poisson}(\omega)$$

$$N_i = K_i + Q_i$$

$$k_i \sim \text{Hypergeometric}(K_i, N_i, n_i)$$

3 | APPLICATIONS

In all the analyses presented below, implementation was done in a Bayesian framework, using program JAGS (Plummer, 2003), called from R (R Core Team 2016) with the package "rjags" (Plummer, 2013). Flat priors were used for all parameters, as follows: (i) a uniform U(0, 1) distribution for probabilities of occupancy (ψ) and site-level true (p_{11}) and false (p_{10}) positive detections; and (ii) a uniform (0, 1000) distribution for the Poisson parameters λ and ω respectively. For each MCMC run, we used a burn-in period of at least 200 iterations to ensure good convergence, which was assessed using the Gelman–Rubin convergence diagnostic (GR). We considered that appropriate convergence was reached when GR < 1.05, and if this threshold was not reached after the first 200 iterations, we prolonged the burn-in period as much as needed. Convergence was usually reached within 200 to 800 iterations. We then used an additional 2,000 MCMC iterations, with two parallel chains, to sample the posterior distribution. The code used for implementation, in BUGS-language programmes, of the three models presented here are provided in Appendix S1.

3.1 | Simulations

Simulated data were used to assess the performance of models *G*, *S1* and *S2* for a range of parameter values, as well as to assess how estimation accuracy of model *G* improved with an increasing proportion of validated data. Data were simulated using a four-step process (Appendix S2) corresponding to the four hierarchical components describing the general model *G*. We generated data, for $I = 70$ and $J = 10$, under 64 different scenarios, corresponding to all combinations of the following parameter values: $\psi = \{0.3, 0.6\}$, $p_{11} = \{0.4, 0.8\}$, $p_{10} = \{0.1, 0.4\}$, $\lambda = \{50, 200\}$, $\omega = \{30, 100\}$ and a total number of validated data $n_{\cdot} = \sum_{i=1}^{I} n_i = \{100, 500\}$. To further investigate how validation effort influences estimation accuracy of ψ, we analysed an additional 14 scenarios of simulated data generated for the following values of the proportion ($n_{\cdot}/N_{\cdot}$): 0.2%, 0.3%, 0.4%, 0.5%, 0.6%, 0.8%, 1%, 2%, 3%, 4%, 5%, 6%, 8% and 10%. For these latter 14 scenarios, the other parameters were fixed to: $\psi = 0.6$, $p_{11} = 0.4$, $p_{10} = 0.4$, $\lambda = 200$, $\omega = 100$. These values were chosen from the worst case scenario identified from the first set of simulations (see Figure 1). In addition, we assessed model performance for cases where very few detections, overall, were expected at each site: $\lambda = \{3,7\}$, $\omega = \{2,5\}$; and for which, many fewer detections were thus validated: $n_{\cdot} = \{5,10,20,50\}$. For these latter scenarios, we also used $\psi = 0.6$, $p_{11} = 0.4$ and $p_{10} = 0.4$. For each scenario, a total of 100 different datasets was simulated and analysed with the models to characterize estimator error.

3.2 | Real datasets

We also assessed the performance of the three estimators with two real datasets consisting of acoustic recordings of calls from pig frogs and green treefrogs. Here, we focus on frog calls, and acknowledge the fact that these are usually less complex than birds calls (e.g. songbirds). Our approach will work similarly with bird calls data, as long as they can be adequately processed to provide a multi-detections/non-detections dataset. For both our frog species, acoustic recordings were collected from 60 sites located in Picayune Strand State Forest in southern Florida (see Walls et al., 2014 for study background), between 04 October 2011 and 16 December 2012. Acoustic recording units were programmed to be active for the first 5 min of every hour (hereafter referred to as a "5-min-periods"), during six consecutive hours at night, for a total of 30 min (5 min per hour × 6 hour) per night. Over the entire recording period (440 days in 2011 and 2012), there was thus a total of 2,640 5-min-periods, representing a potential of up to 220 hours of acoustic recording for each site, with some possible gaps due to temporary ARU failure. After discounting failures, we found that, on average, ARUs actually recorded during 2,032 5-min-periods (*SD* = 417.9), which is equivalent to 169.4 hr (*SD* = 34.8 hr). The range of variation across sites extended from 625 (=52.1 hr) to 2590 5-min-periods (=215.8 hr). Each 5-min-period of acoustic recording produced an individual audio file that was then analysed by a custom created recognizer file using the automated detection algorithm of Song Scope acoustic analysis software (Version 4.1.3, Wildlife Acoustics, Concord, MA).

The automated detection programme can be "tuned" to different sets of sensitivity and specificity, resulting in a trade-off between the number of false negatives and false positives produced by the analysis. This tuning is accomplished by setting different detection thresholds for acoustic signals representing potential detection of the species of interest. The more sensitive the detection threshold, the more likely we are to avoid false negatives and increase false-positive rate; the more specific the detection threshold (i.e. more conservative setting), the higher the rate of false negatives but the lower the false-positive rate (Waddle, Thigpen, & Glorioso, 2009). For our analyses, we chose a conservative setting to minimize false positives. Given the number of recordings available, we were confident that a large number of true detections would still be available, despite the increased rate of false negatives induced by this conservative setting. The modelling approach would still work with a different setting. The raw count data obtained from applying the algorithm consisted of the number of detections, potentially involving both

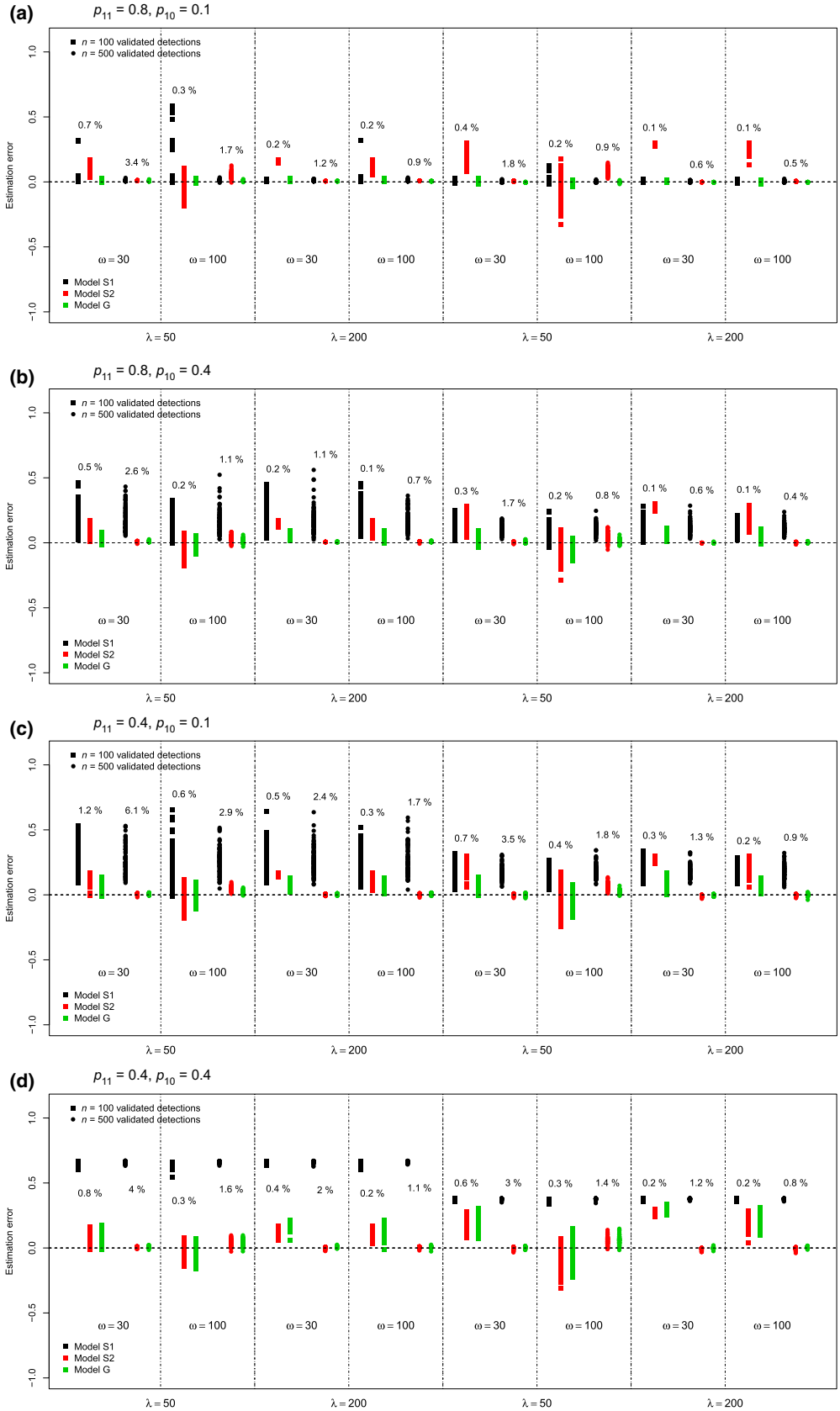


FIGURE 1 Results of model performance assessed from the simulated data analysis. Parameters are defined as follows: λ is the expect number (mean of Poisson distribution) of true detections; ω is the expected number (mean of Poisson distribution) of false positives; *p*11 and *p*10 are site-level probabilities of true and false positives, respectively. The percent values shown on the graph correspond to the proportions of detections validated

true and false positives, within each 5-min-period, available for each site, over the entire sampling period. For the statistical analysis, we extracted raw data from the peak acoustic activity, for each species' dataset, i.e. a subset of the entire recording period. This peak fell between September 09 and October 17 for the pig frog dataset, and between August 18 and October 17 for the green treefrog dataset. Focusing on these limited periods of high acoustic intensity avoided dealing with unnecessarily large datasets, as well as excesses of 0's. It also better ensured that the closure assumption would be valid. Then, we lumped the raw data as unique sampling occasions spanning 9-day periods, providing a total of $J = 8$ and $J = 6$ occasions per site, for the pig frog and the green treefrog datasets respectively. Each occasion j thus corresponded to a total of 270 min of recording (i.e. 9 nights × 30 min per night) and data M_{ij} for a specific site i and occasion j represent all algorithm detections produced during that 9-day time period.

For each dataset, we post-validated *all detections* produced by the algorithm, i.e. $n. = N.$, where $N. = \sum_{i=1}^{I} N_i$ is the total number of detections and $n. = \sum_{i=1}^{I} n_i$ is the total number of detections validated, across all sites. Post-validation was performed directly by an expert herpetologist who listened to all algorithm detections and marked them as "true positive", if the corresponding acoustic fragments contained a true call from the focal species, and "false positive" otherwise. We used information from the 100% validated algorithm detections ($n. = N.$) as the gold standard for determining "true" parameter values, i.e. the best-available estimates from these information-rich datasets. Given the high number of detections produced, we are indeed confident that the 100% validated data provide very accurate estimates of the *sample-at-hand* occupancy probability. Even with a very low detection rate, it is virtually impossible that, at an occupied site, no true call would have ever been recorded during the extended recordings time available for that site. In any case, parameter values provided by this *near-perfect* dataset represent the best estimates that could be achieved by a reduced dataset (i.e. a subset where $n. < N.$) of the whole available sample (i.e. the 100% validated data, $n. = N.$). To assess model performance, models were thus implemented on datasets for which only a fraction ($n./N. < 100\%$) of the validated data were kept. The n_i detections that were kept as validated were randomly selected for each site.

The total number of algorithm detections, including both true and false detections, varied substantially among sites, for both species. For the pig frog, the number of detections per site and occasion ranged from 0 to 158. The total number of detections per site, summed across occasions, ranged from 0 to 661, with a mean of 26.6 ($SD = 104.2$, CV = 3.92) and a median of 3. For the green treefrog dataset, the *per site-occasion* values ranged from 0 to 1,424. The *per site* number of detections ranged from 0 to 2,980, with a mean of 345.1 ($SD = 589.4$, CV = 1.71) and a median of 108.

Most of this variability was due to among-site variation in true positive call frequency (pig frog: CV = 4.34; green treefrog: CV = 2.37). The number of false positives, on the other hand, was more homogeneous across sites (pig frog: CV = 1.28; green treefrog: CV = 1.73). For the pig frog dataset, the *per-site* number of false detections ($M = 3.32$, $SD = 4.25$) was much smaller than the number of true detections ($M = 23.3$, $SD = 100.9$). For the green treefrog data, they were of the same order of magnitude (false detections: $M = 102.6$, $SD = 177.6$; true detections: $M = 242.5$, $SD = 575.4$). Overall, the green treefrog dataset had substantially more detections (both true and false positives) than the pig frog dataset (20,705 vs. 1,595 respectively). We chose these two contrasting examples on purpose, to better assess the performance of our new analytical approach.

4 | RESULTS

4.1 | Simulations

Overall, model *G* provided accurate estimates of occupancy probability, and it always performed at least as well, and often better than, models *S1* and *S2* (Figure 1). Model *S1*, which did not explicitly use the information provided by detection counts, displayed the highest errors of the three models. With this model, the bias in $\hat{\psi}$ dramatically increased with higher *site-level* probability of false positives (i.e. higher p_{10}; see Figure 1d). Moreover, this model induced large imprecision in ψ estimates for many scenarios (Figure 1b,c). Overall, model *S2* tended to perform similar to model *G* when a large sample of validated data ($n. = 500$) was provided, but it became biased and less precise when a small sample of data was validated ($n. = 100$). As for model *S1* and *S2*, the performance of model *G* was negatively affected by increasing values of p_{10} and decreasing values of p_{11} (Figure 1d). However, even in the worst case scenarios (Figure 1d), highly accurate estimates of ψ could be obtained by increasing the number of detections being validated from 100 to 500, which represented a moderate increase in terms of the absolute proportion $n./N.$ (e.g. from 0.2% to about 1%). This trend was confirmed by our targeted investigation of the influence of detection-validation proportions $n./N.$ on estimation accuracy (Figure 2). Indeed, we found that estimates of ψ quickly become unbiased and highly precise as the proportion of data being validated gets close to 1% or higher. We, however, highlight that with the scenarios we used, such small proportions still represent fairly high absolute values of n_i (e.g. $n. \approx 30$ for 0.2%, and $n. \approx 150$ for 1%). When much smaller numbers of total detections ($N.$) are available, higher validation proportions are needed to reach great accuracy (see Figure 3). In these latter cases, however, these small proportions actually represent small numbers of validated detections (e.g. $n. \leq 50$).

4.2 | Real datasets

Model implementation and convergence were successful for both datasets and all three models. From the 100% validated dataset, occupancy probability values were found to be $\psi = 0.27$ for the pig frog data and $\psi = 0.62$ for the green treefrog data.

We first describe results of the analysis of the pig frog data (Figure 4a). We assessed scenarios for the following levels of validation efforts: $n. = \{50, 60, 80, 100, 200, 500\}$, which correspond to proportions lying between 3.1% and 31.3% (see Figure 4a). Model *G* appeared mostly unbiased and provided very precise estimates of ψ when fairly large proportions of detections were validated (e.g. $n. = 200$). With lower validation effort, there was greater imprecision, but the estimates provided by this model were still accurate

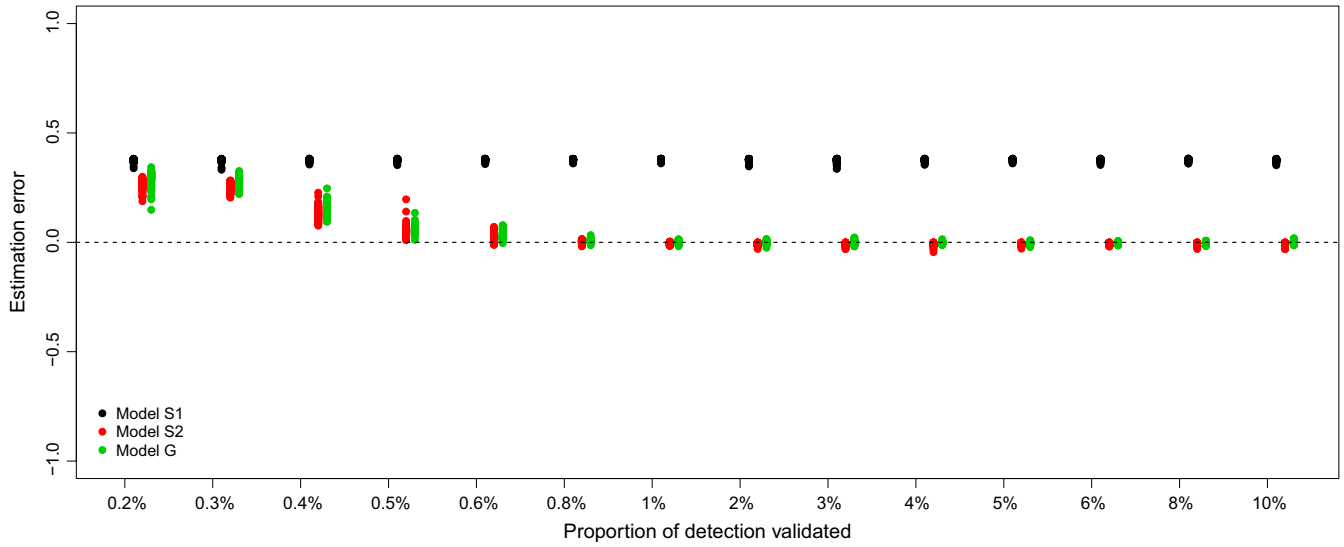


FIGURE 2 Influence of the proportion of detections validated on estimation accuracy

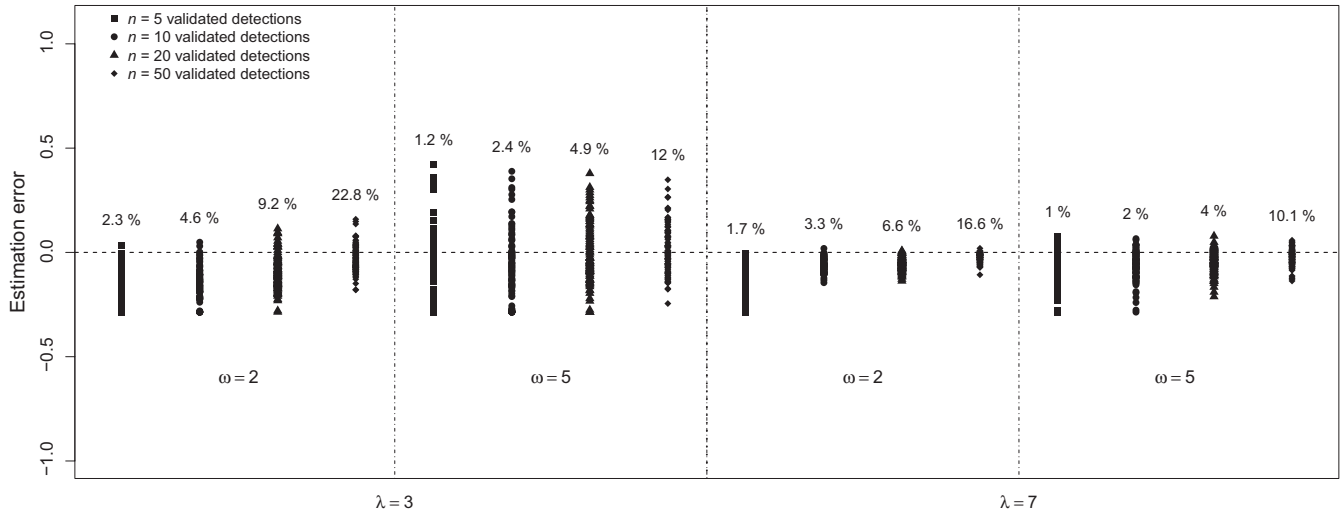


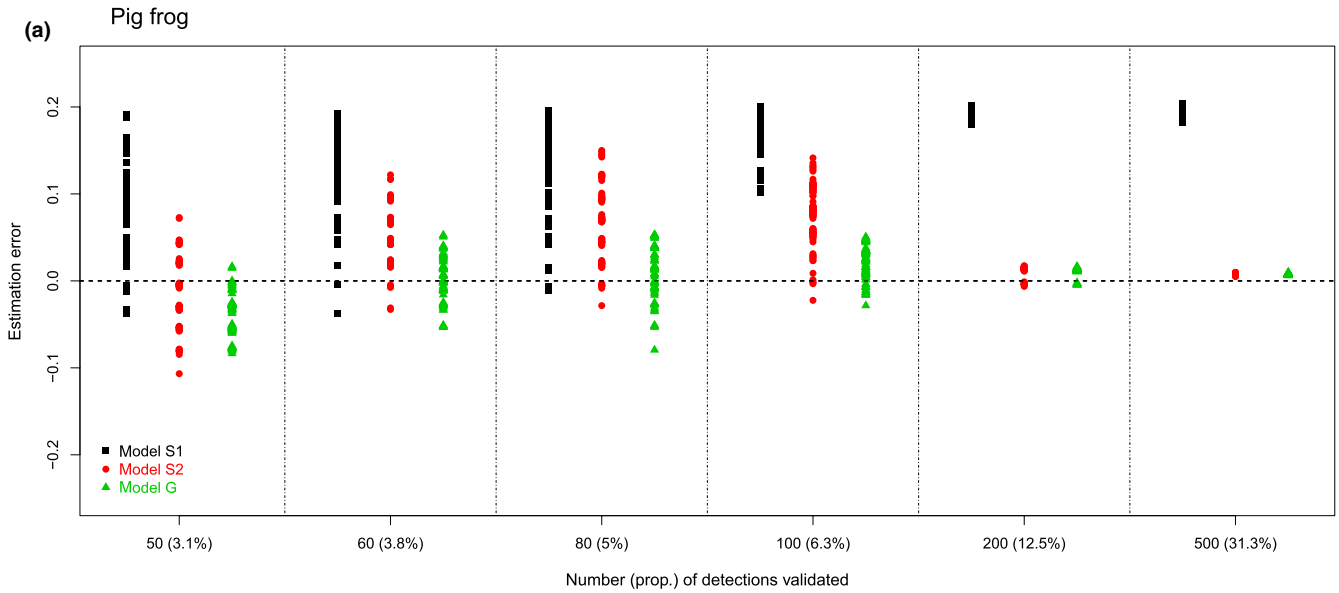
FIGURE 3 Estimation accuracy of model *G* for cases where fewer detections are expected (i.e. small values of λ and ω)

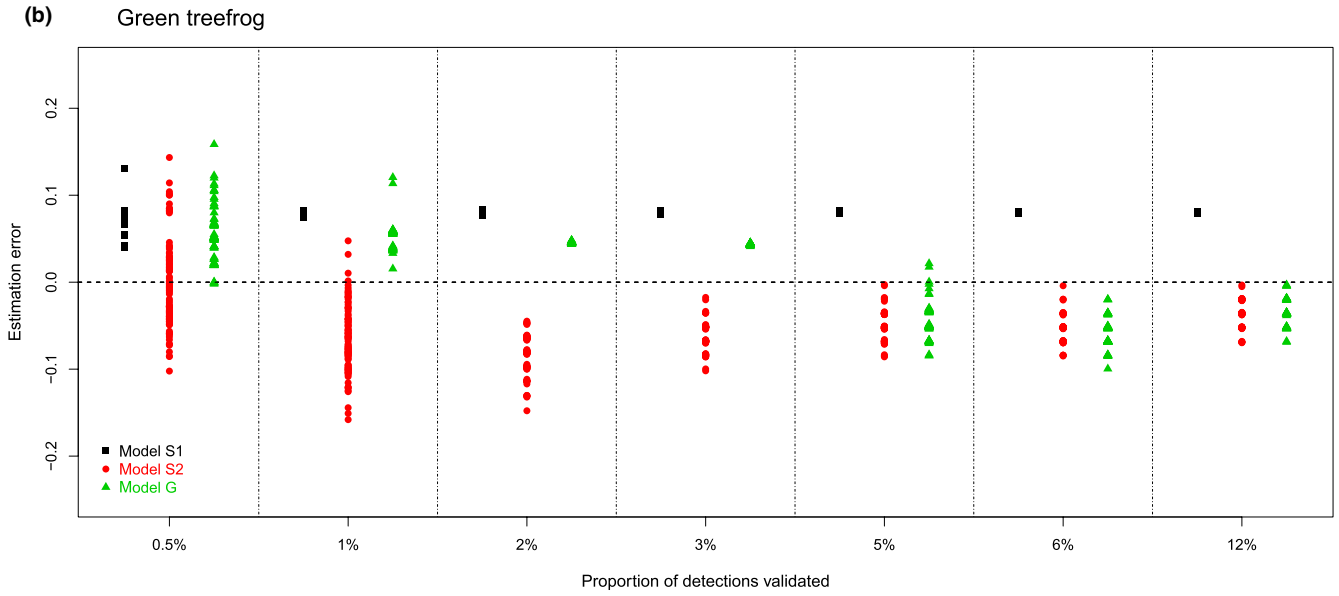
enough to be useful and trustworthy. Model *S2* also provided very accurate estimates at large levels of validation effort, but it did not perform nearly as well in the lower effort range. Especially, when about 100 or fewer detections were validated, estimates of ψ were slightly positively biased. Model *S1* performed poorly at all levels of validation efforts, displaying a severe positive bias, constant across n_i values, and large imprecision at low n_i values.

For the analysis of the green treefrog data, which had over 20,000 detections total, we assessed scenarios for the following range of proportions *n./N.* = {0.5%, 1%, 2%, 3%, 5%, 6%, 12%}. A first analysis of the entire treefrog dataset revealed that estimation accuracy from model *G* was not as good as with the pig frog analysis (Figure 4b). As in the simulation analyses, a positive bias and substantial imprecision were observed for all validation efforts lower than 2% (i.e. *n.* < 400). At validation effort ≥5%, the estimate became mostly unbiased, but it bore some imprecision. However, after we removed five outlier sites (see Section 5), estimation greatly improved (Figure 4c). Model *S2* was unbiased, but imprecise, at low values of validation efforts (<1%). Within this range, it performed slightly better than model *G*. At the highest values of validation effort assessed, model *G* and *S2* performed equally well (Figure 4b,c). As for the other analysis, model *S1* performed poorly, overall, displaying positive biases, as well as larger imprecision at low *n.* values.

5 | DISCUSSION

We present methods to efficiently estimate site occupancy probability when a species is monitored intensively, such that at any site-occasion, multiple detections might be available. This is the case for automated acoustic recorders, which motivated our interest and provided the case study we use to illustrate the model. As discussed by previous authors (Brandes, 2008; Lammers et al., 2008), acoustic recorders provide such amounts of data that one usually must resort to automated algorithms


(a)
Pig frog
Estimation error
0.2
0.1
0.0
−0.1
−0.2
Model S1
Model S2
Model G
50 (3.1%)
60 (3.8%)
80 (5%)
100 (6.6%)
200 (12.5%)
500 (31.3%)
Number (prop.) of detections validated


(b)
Green treefrog
Estimation error
0.2
0.1
0.0
−0.1
−0.2
Model S1
Model S2
Model G
0.5%
1%
2%
3%
5%
6%
12%
Proportion of detections validated

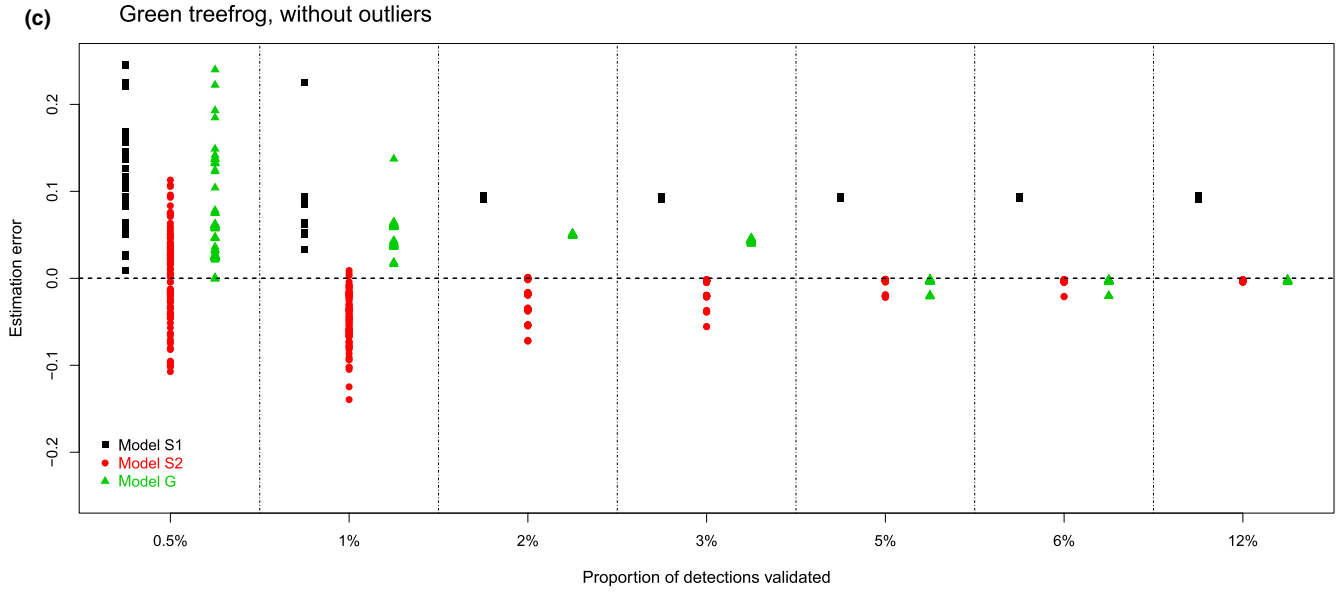

(c)
Green treefrog, without outliers
Estimation error
0.2
0.1
0.0
−0.1
−0.2
Model S1
Model S2
Model G
0.5%
1%
2%
3%
5%
6%
12%
Proportion of detections validated

FIGURE 4 Results from analyses of (a) the pig frog data, (b) the entire green treefrog dataset and (c) the green treefrog dataset after removing five outlier sites (see Section 5). For each scenario considered (i.e. a given proportion of validated detections), individual estimate errors, obtained from 100-time repeated analyses, are shown for each of the three models

to efficiently extract detections over the entire recording time frame. Because automated recognition programmes are not perfect, they generate two types of observation errors: (1) some of the species sounds (e.g. calls) present on the acoustic record will not be identified by the programme, thus introducing false negatives, in addition to false negatives resulting from the species not calling during recording time; and (2) some other sounds (e.g. from another species or background noise) might be confounded with those of the focal species and identified as detections, introducing false positives in the data. Our general model (*G*), presented in a hierarchical framework, builds upon pre-existing occupancy models (Chambert, Miller et al., 2015; Miller et al., 2011, 2013) to provide an efficient way of accounting for these two types of observation error.

The novelty of this approach is that we fully use the information provided by the multiplicity of detections using Poisson processes instead of simply collapsing the information to a traditional binary response (0/1) of detection/non-detection for any site-occasion considered. This provides key additional information to better disentangle false positives from true detections. In addition, we include a process for incorporating validated observations, where a subset of detections can be verified, *post hoc*, as true or false positives. The validation mechanism, based on random selection of a subset of the data, is explicitly modelled with a hypergeometric process. In addition to evaluating this general model *G*, we assessed performance of two alternative models, *S1* and *S2*, which can be written as special cases of our general model formulation. First, model *S1* is equivalent to the binomial mixture model of Royle and Link (2006), with the addition of known statuses provided for a subset of sites for which true detections were confirmed through the partial validation process. This latter model thus ignores the quantitative information provided by detection counts. On the other hand, model *S2* explicitly models detection counts through Poisson processes, but it does not contain any binomial component for the occurrence of false negatives. Its ability to account for the latter type of observation error is thus much more limited.

Simulation results revealed that, overall, model *G* is unbiased and can provide very precise estimates for occupancy probabilities of a species. Of the three models assessed, model *G* consistently performed the best. Model *S2*, which is slightly simpler to implement, can be a good alternative, especially when it is not too costly to validate additional detections (e.g. 500 vs. 100). Alternatively, using the much simpler model, *S1*, which ignores the information from detection counts, only performs well when p_{11} is high and p_{10} is low.

Results for the pig frog data were consistent with the simulation results. Indeed, model *G* performed well, and better than the two alternative models. Model *S2* displayed large bias, except at high levels of validation efforts, where it performed similar to model *G*. Model *S1* provided unreliable estimates (severely biased and/or very imprecise). With the green treefrog data, model *G* and *S2* did not perform quite as well. This difference of performance was explained by a higher level of among-site heterogeneity in the green treefrog dataset (variance-to-mean ratio in per-site detection counts: *c.* 400 for the pig frog dataset vs. *c.* 1,000 for the green treefrog dataset). More specifically, further investigations led us to identify five sites for which the latent occupancy status was consistently underestimated. These five sites were clearly outliers, as they all had very few true detections (<12) and high number of false positives (between 132 and 761). This contrasts with other occupied sites, which had, on average, about 454 true detections and only 86 false positives. This represents a 500 fold difference, in terms of true-positives-to-false-positives ratio (TP/FP), between these five outlier sites (TP/FP = 0.17) and other occupied sites (TP/FP = 9.0). The relative abundance of true positives to false positives is thus an important factor to consider, and substantial heterogeneity in this ratio can cause estimation problems. This was confirmed by a follow-up analysis, in which we removed these five outlier sites. Estimates became much more accurate (Figure 4c) and followed the same pattern as observed for the simulated data (Figure 2) and the pig frog analysis (Figure 4a).

To improve estimation in the presence of false-positive detections, it is often desirable to know the true status of a subset of detections (Chambert, Miller, et al., 2015; Miller et al., 2011). Here, this was accomplished by the post-validation, by an expert herpetologist, of a subsample *n.* of algorithm detections. Although we are confident about the quality of the expertise, we cannot always rule out the possibility of misidentification by the expert. If this is suspected to occur, steps could be taken to estimate the impact of validation errors on the performance metrics. Our simulations revealed that a small number of detections, relative to the numbers of sites, need to be validated to obtain accurate occupancy estimates. For scenarios with large number of detections at each sites ($\lambda = 200$, $\omega = 100$), only *c.* 1% (i.e. *n.* ≈ 150 validated detections) was needed (see Figure 2). In the presence of many fewer detections (e.g. $\lambda = 7$, $\omega = 5$), higher proportions were required, but these represented very few detections to validate (e.g. *n.* ≈ 20–50; see Figure 3). When these number are scaled to the number of sites ($I = 70$), we conclude that validation of *c.* 0.5–3 detections per site is sufficient to reach excellent estimator accuracy, suggesting that partial verification can be a very efficient method to improve accuracy of estimates.

Our goal was to lay the foundation for an innovative approach of analysing acoustic and similar information-rich species detection data, leveraging the availability of multiple detections to better account for false positives. Exciting opportunities to further develop this modelling approach exist. First, to better understand the dynamic processes that lead to changes in species distributions, it will be essential to extend this model to accommodate multi-year occupancy dynamics. Such dynamic occupancy models have become commonplace for estimating site extinction and colonization rates of metapopulations (Bailey, MacKenzie, & Nichols, 2014; MacKenzie et al., 2003). Results from Miller et al. (2015) suggest that estimators accounting for false positives may be more efficient in dynamic models where probability of occurrence in previous and future years increases identifiability of identification error types. Second, the development of multistate extensions (Nichols, Hines,

MacKenzie, Seamans, & Gutiérrez, 2007) of models *S2* and *G* will also allow investigating new ecological questions (Royle, 2004). Indeed, the use of several occupancy states, based on call intensity and frequency, could provide insights into local species activity, habitat use, or could be used as indirect metrics of local abundance or density, as was done by Royle (2004). Third, the extra information provided by multiple detections could also be used towards more direct and quantitative estimation of activity metrics or abundance indexes. In our current model, we use a (Zero-Truncated) Poisson process to model the intensity of true calls in our data, which confounds two different processes: (1) the intensity of calls truly emitted by the species during recording time and (2) the probability that any of these available calls is detected by the algorithm. These two components could be disentangled and modelled separately, in a hierarchical way, providing direct estimation of the true, latent intensity of calls emitted. Such estimates of call abundance could not be directly interpreted as animal abundance (Corn, Muths, Kissel, & Scherer, 2011), but they can nevertheless provide a useful metric of species activity, and could be used for the development of approximate indices of local abundance. Such indices could provide important insights for species-habitat relationships at a finer scale than occupancy metrics (e.g. Knutson et al., 1999; Price, Browne, & Dorcas, 2012). They could also be used in dynamic models to investigate species phenology along the season or responses to habitat changes (Chambert, Kendall, et al., 2015). Finally, as automated detection technologies usually provide detections for many different species, the modelling framework presented here could be expanded to investigate questions related to occupancy, species richness or other derived metrics, for entire communities (Kéry, Royle, Plattner, & Dorazio, 2009).

The development of new species detection technologies, which have rapidly expanded in the last two decades, has great promise for animal monitoring programmes and ecological studies. However, it is important to develop adequate statistical and modelling approach that are tailored to deal with new sources of observation errors and provide accurate parameter estimates (Kéry & Schmidt, 2008; Pollock et al., 2002). In addition to statistical developments, it is also crucial to ensure that appropriate sampling designs are being proposed and implemented for monitoring studies relying on such technologies (O'Connell et al., 2010; Williams, Nichols, & Conroy, 2002). Technology and statistical methods will never compensate for biases induced by haphazard study designs, nor can they replace the need for any investigator to ask *why* he/she is implementing a given monitoring programme and *what* metric he/she should seek to best address his/her question (Guillera-Arroita et al., 2015). If appropriately used, new detection technologies, in combination with appropriate analysis methods such as the ones presented and discussed in this paper, will expand the boundaries of ecological investigation in many under-studied systems.

ACKNOWLEDGEMENTS

This work was supported by the United States Geological Survey's Greater Everglades Priority Ecosystem Science programme (USGS GEPES) and the Amphibian Research and Monitoring Initiative (USGS ARMI). J. Barichivich, I. Bartoszek, M. Brown, B. Glorioso, and J. Hefner assisted in the collection of ARU data, and J. Hefner performed validation of frog call identifications. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This is contribution 595 of USGS ARMI.

AUTHORS' CONTRIBUTIONS

T.C., D.M. and J.N. conceived and developed the methodology; J.H.W. and S.W. conceived the field sampling protocols and collected the data; T.C. analysed the data; T.C. and D.M. led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA ACCESSIBILITY

All the data and code of this article are available in a Dryad Digital Repository https://doi.org/10.5061/dryad.949nc (Chambert, Waddle, Miller, Walls, & Nichols, 2017). Data files are also provided in Supporting Information and available from Waddle, Hefner, and Walls (2017).

ORCID

Thierry Chambert http://orcid.org/0000-0002-9450-9080

REFERENCES

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., & Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, *4*, 206–214.

Acevedo, M. A., & Villanueva-Rivera, L. J. (2006). Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. *Wildlife Society Bulletin*, *34*, 211–214.

Bailey, L. L., MacKenzie, D. I., & Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, *5*, 1269–1279.

Bardeli, R., Wolff, D., Kurth, F., Koch, M., Tauchert, K.-H., & Frommolt, K.-H. (2010). Detecting bird sounds in a complex acoustic environment and application to bioacoustic monitoring. *Pattern Recognition Letters*, *31*, 1524–1534.

Blumstein, D. T., Mennill, D. J., Clemins, P., Girod, L., Yao, K., Patricelli, G., … Kirschel, A. N. G. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *Journal of Applied Ecology*, *48*, 758–767.

Brandes, T. S. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, *18*, S163–S173.

Chambert, T., Kendall, W. L., Hines, J. E., Nichols, J. D., Pedrini, P., Waddle, J. H., … Tenan, S. (2015). Testing hypotheses on distribution shifts and changes in phenology of imperfectly detectable species. *Methods in Ecology and Evolution*, *6*, 638–647.

Chambert, T., Miller, D. A. W., & Nichols, J. D. (2015). Modelling false positive detections in species occurrence data under different study designs. *Ecology*, *96*, 332–339.

Chambert, T., Waddle, J. H., Miller, D. A. W., Walls, S. C., & Nichols, J. D. (2017). Data from: A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.949nc.

Clark, C. W., & Clapham, P. J. (2004). Acoustic monitoring on a humpback whale (*Megaptera novaeangliae*) feeding ground shows continual

singing into late spring. *Proceedings of the Royal Society of London B: Biological Sciences*, *271*, 1051–1057.
Corn, P. S., Muths, E., Kissel, A. M., & Scherer, R. D. (2011). Breeding chorus indices are weakly related to estimated abundance of boreal chorus frogs. *Copeia*, *2011*, 365–371.
Crouch, W. B., & Paton, P. W. C. (2002). Assessing the use of call surveys to monitor breeding anurans in Rhode Island. *Journal of Herpetology*, *36*, 185–192.
Dorcas, M. E., Price, S. J., Walls, S. C., & Barichivich, W. J. (2010). Auditory monitoring of anuran populations. In C. K. J. Dodd (Ed.), *Amphibian ecology and conservation: A hand book of techniques* (pp. 281–298). Oxford, UK: Oxford University Press.
Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., … Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, *24*, 276–292.
Haselmayer, J., & Quinn, J. S. (2000). A comparison of point counts and sound recording as bird survey methods in Amazonian Southeast Peru. *The Condor*, *102*, 887–893.
Kéry, M., Royle, J. A., Plattner, M., & Dorazio, R. M. (2009). Species richness and occupancy estimation in communities subject to temporary emigration. *Ecology*, *90*, 1279–1290.
Kéry, M., & Schmidt, B. (2008). Imperfect detection and its consequences for monitoring for conservation. *Community Ecology*, *9*, 207–216.
Knutson, M. G., Sauer, J. R., Olsen, D. A., Mossman, M. J., Hemesath, L. M., & Lannoo, M. J. (1999). Effects of Landscape Composition and Wetland Fragmentation on Frog and Toad Abundance and Species Richness in Iowa and Wisconsin, USA. *Conservation Biology*, *13*, 1437–1446.
Lammers, M. O., Brainard, R. E., Au, W. W. L., Mooney, T. A., & Wong, K. B. (2008). An ecological acoustic recorder (EAR) for long-term monitoring of biological and anthropogenic sounds on coral reefs and other marine habitats. *The Journal of the Acoustical Society of America*, *123*, 1720–1728.
MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, *84*, 2200–2207.
MacSwiney, G., Cristina, M., Clarke, F. M., & Racey, P. A. (2008). What you see is not what you get: The role of ultrasonic detectors in increasing inventory completeness in Neotropical bat assemblages. *Journal of Applied Ecology*, *45*, 1364–1371.
McClintock, B. T., Bailey, L. L., Pollock, K. H., & Simons, T. R. (2010). Unmodelled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, *91*, 2446–2454.
Miller, D. A. W., Bailey, L. L., Campbell Grant, E. H., McClintock, B. T., Weir, L. A., & Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: A test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, *6*, 557–565.
Miller, D. A. W., Brehme, C. S., Hines, J. E., Nichols, J. D., & Fisher, R. N. (2012). Joint estimation of habitat dynamics and species interactions: Disturbance reduces co-occurrence of non-native predators with an endangered toad. *The Journal of Animal Ecology*, *81*, 1288–1297.
Miller, D. A. W., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., & Mitchell, M. S. (2013). Determining occurrence dynamics when false positives occur: Estimating the range dynamics of wolves from public survey data. *PLoS ONE*, *8*, e65808.
Miller, D. A. W., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, *92*, 1422–1428.
Miller, D. A. W., Weir, L. A., Mcclintock, B. T., Grant, E. H. C., Bailey, L. L., & Simons, T. R. (2012). Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications: A Publication of the Ecological Society of America*, *22*, 1665–1674.
Nichols, J. D., Hines, J. E., MacKenzie, D. I., Seamans, M. E., & Gutiérrez, R. J. (2007). Occupancy estimation with multiple states and state uncertainty. *Ecology*, *88*, 1395–1400.
O'Connell, A. F., Nichols, J. D., & Karanth, K. U. (2010). *Camera traps in animal ecology: Methods and analyses*. New York: Springer.
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Proceedings of the 3rd international workshop on distributed statistical computing, p. 125. Technische Universit at Wien, Wien, Austria.
Plummer, M. (2013). *rjags: Bayesian graphical models using MCMC*. Retrieved from http://cran.r-project.org/package=rjags .
Pollock, K. H., Nichols, J. D., Simons, T. R., Farnsworth, G. L., Bailey, L. L., & Sauer, J. R. (2002). Large scale wildlife monitoring studies: Statistical methods for design and analysis. *Environmetrics*, *13*, 105–119.
Price, S. J., Browne, R. A., & Dorcas, M. E. (2012). Evaluating the effects of urbanisation on salamander abundances using a before-after control-impact design. *Freshwater Biology*, *57*, 193–203.
R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/
Rempel, R. S., Hobson, K. A., Holborn, G., Van Wilgenburg, S. L., & Elliott, J. (2005). Bioacoustic monitoring of forest songbirds: Interpreter variability and effects of configuration and digital processing methods in the laboratory. *Journal of Field Ornithology*, *76*, 1–11.
Royle, J. A. (2004). Modelling abundance index data from anuran calling surveys. *Conservation Biology*, *18*, 1378–1385.
Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, *87*, 835–841.
Simons, T. R., Alldredge, M. W., Pollock, K. H., & Wettroth, J. M. (2007). Experimental analysis of the auditory detection process on avian point counts. *The Auk*, *124*, 986–999.
Waddle, J. H., Hefner, J. M., & Walls, S. C. (2017). Computer automated frog vocalization results from Picayune Strand State Forest, Florida 2011-2012. *U.S. Geological Survey data release*, https://doi.org/10.5066/f7mp51h4
Waddle, J. H., Thigpen, T. F., & Glorioso, B. M. (2009). Efficacy of automatic vocalization recognition software for anuran monitoring. *Herpetological Conservation and Biology*, *4*, 384–388.
Walls, S. C., Waddle, J. H., Barichivich, W. J., Bartoszek, I. A., Brown, M. E., Hefner, J. M., & Schuman, M. J. (2014). Anuran site occupancy and species richness as tools for evaluating restoration of a hydrologically-modified landscape. *Wetlands Ecology and Management*, *22*, 625–639.
Williams, B. K., Nichols, J. D., & Conroy, M. J. (2002). *Analysis and management of animal populations*. New York, NY: Academic Press.
Wrege, P. H., Rowland, E. D., Keen, S., & Shiu, Y. (2017). Acoustic monitoring for conservation in tropical forests: Examples from forest elephants. *Methods in Ecology and Evolution*, *8*, 1292–1301.
Zwart, M. C., Baker, A., McGowan, P. J. K., & Whittingham, M. J. (2014). The use of automated bioacoustic recorders to replace human wildlife surveys: An example using nightjars. *PLoS ONE*, *9*, e102770.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Chambert T, Waddle JH, Miller DAW, Walls SC, Nichols JD. A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods Ecol Evol*. 2018;9:560–570.
https://doi.org/10.1111/2041-210X.12910